Data Extraction and Mining of Reddit User Data

L Pranau Kumar (l.pranaukumar2015@vit.ac.in)

School of Computer Science Engineering

VIT University, Vellore

Abstract

Reddit is a popular user driven forum that consists of many communities called subreddits dedicated to discussing a pre-defined topic. Users can submit original content or links to other sites and have discussions. This is a unique social network because the emphasis is on the community rather than a single user. In this project, we scraped Reddit for varied posts and ordered them into a single data set. We then examined the structure of this data set. Through this data set we analyzed what the current topics of discussions were about, what the perceived opinions of the users were about the various topics. Analyzed data regarding discussions about political trends, events and other policy decisions. We scraped the data using APIs provided by Reddit and applied data mining and sentiment analysis techniques on the scraped data. Through this analysis, we gained a clear insight into how users of such sites behave and interact with one another while also gaining insights into their opinions and biases towards various issues.

Keywords- *Web Mining, Reddit, Sentiment Analysis, Network Visualization, Python, R*

I. Introduction

Reddit is a popular user-driven forum that consists of many communities called subreddits dedicated to discussing a pre-defined topic. Users can submit original content or links to other sites and have discussions. This is a unique web platform because the emphasis is on the community rather than a single user. In this project, we propose to scrape Reddit for varied posts and order them into a single dataset. We then propose to examine the structure of this data set. Through this data set, we propose to analyze what the current topics of discussions are about, what the perceived opinions of the users are about the various topics. Analyze data regarding discussions about political trends, events, and other policy decisions. We propose to scrape the data using APIs provided by Reddit and apply data mining and sentiment analysis techniques on the scraped data. Through this analysis, we expect to gain a clear insight into how users of such sites behave and interact with one another while also gaining insights into their opinions and biases towards various issues.

Reddit is a social news aggregation and discussion website with large communities of people across different categories. This project will focus on mining/extracting data efficiently from popular subreddits and analyzing that data to form meaningful conclusions about the usage of the website.

Sentiment analysis or opinion mining is the computational study of people's opinions, sentiments, attitudes, and emotions expressed in written language. It is one of the most active research areas in natural language processing and text mining in recent years.

The growing importance of sentiment analysis coincides with the growth of social media such as reviews, forum discussions, blogs, micro-blogs, Twitter, and web platforms. For the first time in human history, we now have a huge volume of opinionated data recorded in digital form for analysis. Sentiment analysis systems are being applied in almost every business and social domain because opinions are central to almost all human activities and are key influencers of our behaviors. Our beliefs and perceptions of reality, and the choices we make, are largely conditioned on how others see and evaluate the world.

The task is to study the dataset and prepare a proposal of what knowledge we

plan to extract from this dataset using a data-mining technique. The proposal must then be implemented along with a report of the related works, analysis, techniques, methodology, evaluation, and results.

Creating, placing, and characterizing social media comments and reactions is a challenging problem. This is particularly true for reddit.com, a highly trafficked social media website with thousands of posts per day. Each post has an associated comment thread, and users of Reddit can vote the comments up or down, generating a net score, or "Karma," for each comment. Users aspire to collect this "Karma," and these comments build the community on Reddit. When reading the content of the comment thread, however, it is often unclear why some comments succeed and receive high Karma while other comments lose Karma. This project explores better insight into Reddit communities through the sentiment analysis of Reddit comments and on Reddit comments to characterize the voting patterns of Reddit users and determines the Karma strength of Reddit comments through identifying comments with positive and negative Karma.

While there has been a fair amount of research on how sentiments are expressed in genres such as online reviews and news articles, how sentiments are expressed given the informal language and message-length constraints of microblogging has been much less studied. Features such as automatic part-of-speech tags and resources such as sentiment lexicons have proved useful for sentiment analysis in other domains, but will they also prove useful for sentiment analysis in Reddit? In this project, we begin to investigate this question. Another challenge of microblogging is the incredible breadth of topic that is covered. It is not an exaggeration to say that people post about anything and everything. Therefore, to be able to build systems to our reddit sentiment about any given topic, we need a method for quickly identifying data that can be used for training. In this project, we explore one method for building such data: using different hashtags (e.g., #bestfeeling, #epicfail, #news) to identify positive, negative, and neutral tweets to use for training three-way sentiment classifiers.

The growing expansion of contents, placed on the Web, provides a huge collection of textual resources. People share their experiences, opinions or simply talk just about whatever concerns them online. A large amount of available data attracts system developers, studying on automatic mining and analysis. In this paper, the primary and underlying idea is that the fact of knowing how people feel about certain topics can be considered as a classification task.

People's feelings can be positive, negative or neutral. A sentiment is often represented in subtle or complex ways in a text. An online user can use a diverse range of other techniques to express his or her emotions. Apart from that, s/he may mix objective and subjective information about a certain topic. On top of that, data gathered from the World Wide Web often contain a lot of noise. Indeed, the task of automatic sentiment recognition in the online text becomes more difficult for all the aforementioned reasons.

II. Objective

Social Media platforms have become quintessential for user-generated content and consumer opinions. The Reddit API and code are open sourced so Reddit's comments can be readily scraped. To expand on that recent reviews have been given by using framework tools like python, R, PRAW, matplotlib and NLTK. The growing expansion of contents, placed on the Web, provides a huge collection of textual resources. People share their experiences, opinions or simply talk just about whatever concerns them online. A large amount of available data attracts system developers, studying on automatic mining and analysis.

Sentiment analysis is becoming a popular study these days, mainly because of the fact that social networking sites include online users who are free to express their thoughts, feelings, and impressions concerning a specific topic. In fact, nowadays, any kind of marketing business is currently immersing to the new trends of businesses. Apart from written surveys, the companies also extend their customer satisfaction analysis through the web, in order to gather a large amount of data.

The specific objectives of the recent study are:

- To develop our own corpus through a Reddit application.

- To properly train the system to accept inputs in the form of status updates from the corpus, disregarding updates that do not contain words or emojis.
- The ability of the system to classify the polarity of an opinion per status update basis, during the testing phase,

2.1 Opinion Mining

The ideal opinion-mining tool would be to process a set of search results for a given item, generating a list of product attributes (quality features, etc.) and aggregating opinions about each of them (poor, mixed, good). However, the term has recently also been interpreted more broadly to include many different types of analysis of the evaluative text. In general, opinions can be expressed on anything, e.g., a product, a service, a topic, an individual, an organization, or an event. The general term object is used to denote the entity that has been commented on. Thus, object O can be defined as an entity which can be a product, topic, person, event, or organization. It is associated with a pair, *O: (T, A)*, where T is a hierarchy or taxonomy of components (or parts) and sub-components of O, and A is a set of attributes of O. Each component has its own set of sub-components and attributes. The word features are used to represent both components and attributes. For an evaluative document D, opinion passage on a feature f of the object O evaluated in D is a group of consecutive sentences in D that expresses a positive or negative opinion on f. Research on opinion mining or sentiment analysis started with identifying opinion (or sentiment) bearing words, e.g., great, amazing, wonderful, bad, and poor. Many researchers have worked on mining such words and identifying their semantic orientations or polarity determination (i.e., positive, negative and neutral). In [12], the authors identified several linguistic rules that can be exploited to identify opinion words and their orientations from a large corpus. This method has been applied, extended and improved in [7, 15, and 22]. In [13 and 16], a bootstrapping approach is proposed, which uses a small set of given seed opinion words to find their synonyms and antonyms in WordNet.

2.2 Sentiment classification

The history of the phrase sentiment analysis parallels that of ―opinion mining‖ in certain respects. The term―sentimental used in reference to the automatic analysis of evaluative text and tracking of the predictive judgments that appear in 2001 paper by Das and Chen. Subsequently, this concept was adopted and enhanced by Turney and Pang in 2005. In the following year, the concept was carried on by Nasukawa & Yi and Yi. These events together may explain the popularity of ―sentiment analysis‖ among communities self-identified as focused on NLP.

In particular, sentiment analysis on online reviews has become a hot research field. Studies on sentiment analysis mainly focus on framework and lexicon construction, feature extraction, and polarity determination.

This review conducts an overall survey of the three major research fields in sentiment analysis: framework, feature extraction and sentiment analysis, making a summary and analysis of the present development, and giving a detailed introduction of its application in business and Blogs. Despite the current immaturity of related research, sentiment analysis of online review has taken its position as an emerging research frontline, which takes advantage of the achievements in many areas, such as text mining, natural language processing, web mining, and machine learning. But the related research did not take place until recently, and semantic parsing and understanding exhibit high complexity, the overall research in this field being in its infantry. Still a lot of problems need further exploration and solution as follows:

(1) Insufficient empirical language data and platform. So far no experimental platform has been released for public use and cast widespread influence; experimental public corpora, especially marked corpora are relatively sparse. The available corpora for English sentiment analysis are MPQA news corpus and the movie critic corpus by Pang. They are confined with limited arrange of tasks and under-satisfied corpus quality. For lack of an open experiment platform and standardized benchmark, it is difficult to assess the effectiveness of all the methods.

(2) No breakthrough in textual sentiment analysis. Deep sentiment analysis inevitably involves semantic parsing, and sentiment transfers in texts moreover, so there is not much progress in deep-structure semantic sentiment analysis, and in textual sentiment analysis.

(3) No research on the commercial value of online product reviews. Online reviews cast a profound influence on the purchase behaviour of consumers.

III. Related Works

[1] While using neural networks has already been applied to analyzing Reddit comments before, there has been past work using Reddit's data to analyze the website. Those contributors sought to evaluate Reddit submissions based on the title of the submission. However, the report explained how the title, submission times, and community choices of image submissions affect the success of the content by investigating resubmitted images on Reddit. The language model used include modeling good and bad words, LDA, parts of speech tagging, length of the title, sentiment analysis. Using this language model they were able to account for how the success of each submission was influenced by its title, and how well that title was targeted to the community. Although using language models is an important factor to the success of the post, the quality of the content, submission time, and the community contributed greatly to the success of the post. [1]

Even we have this kind of analysis for other websites like Twitter, Facebook, Google+ and others.

[2] There have been many papers written on sentiment analysis for the domain of blogs and product reviews. (Pang and Lee 2008) gives a survey of sentiment analysis. Researchers have also analysed the brand impact of microblogging (Jansen). Overall, text classification using machine learning is a well-studied field (Manning and Schuetze 1999). (Pang and Lee 2002) researched the effects of various machine learning techniques (Naive Bayes (NB), Maximum Entropy (ME), and Support Vector Machines (SVM) in the specific domain of movie reviews. They were able to achieve an accuracy of 82.9% using SVM and a unigram model. Researchers have also worked on detecting sentiment in text. (Turner 2002) presents a simple algorithm, called semantic orientation, for detecting sentiment. (Pang and Lee 2004) present a hierarchical scheme in which text is first classified as containing sentiment, and then classified as positive or negative. [2]

Work (Read, 2005) has been done in using emoticons as labels for positive and sentiment. This is very relevant to Twitter because many users have emoticons in their tweets.

Sentiment analysis has been handled as a Natural Language Processing task at many levels of granularity. Starting from being a document level classification task, it has been handled at the sentence level and more recently at the phrase level. There are many established methods for sentiment analysis at the sentence and paragraph level.

In (T. Mullen, N. Collier, 2004), the authors discussed the application of support vector machines in sentiment analysis with the diverse information source.

In (B. Pang, L. Lee, 2004), the authors applied minimum cuts in graphs to extract the subjective portion of texts they were studying and used machine learning methods to perform sentiment analysis on those snippets of texts only.

In (T. Wilson, J. Wiebe, P. Hoffman, 2005), the authors discussed categorizing texts into polar and neutral first before determining whether a positive or negative sentiment is expressed through the text.

However, in (N. Godbole, M. Srinivasaiah, and S. Skiena, 2007), the authors operate on the premise that little neutrality exists in online texts.

In (N. Godbole, M. Srinivasaiah, and S. Skiena, 2007), the authors developed techniques that algorithmically identify large number (hundreds) of adjectives, each with an assigned score of polarity, from around a dozen of seed adjectives. Their methods expand two clusters of adjectives (positive and negative word groups) by recursively querying the synonyms and antonyms from WordNet. Since recursive search quickly connects words from the two clusters, they implemented several precautionary

measures such as assigning weights which decrease exponentially as the number of hops increases. This confirms that the algorithm-generated adjectives are highly accurate by comparing them to the results of manually picked word lists. It is worth pointing out that this work uses Lydia as the backbone to process a large amount of news and blogs.

IV. Frameworks Used

4.1 Python

Python is a high-level programming language for general-purpose program writing, created by Guido van Rossum and was released in 1991. An interpreted language, Python has a design idea that stresses code readability (especially using whitespace indentation to bound code blocks rather than curly brackets or keywords), and a grammar that allows programmers to express concepts in less lines of code than might be used in languages such as Java or C++. The language platform provides constructs planned to permit writing clear line-ups on both a small and large scale.

Python structures a dynamic type system, programmed memory management; support multiple programming paradigms, including object-oriented (OOP), imperative, functional programming, and procedural styles. It has huge and comprehensive standard libraries.

Python interpreters are available for many operating systems (OS); enabling Python code to run on a wide range of systems.

4.2 PRAW

PRAW, a shortening for 'Python Reddit API Wrapper', is a python package that enables for simple access to Reddit's API. PRAW, a shortening for "Python Reddit API Wrapper", is a python package that permits for simple access to Reddit's API. PRAW targets to be easy to use and within follows all of Reddit's API rules. PRAW has enabled the effective data mining and scraping of Reddit in a licensed way so that the bots do no adversely affect the site and the developers can work on their projects for analysis and sentiments extraction.

4.3 Matplotlib

Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK+. There is also a procedural "pylab" interface based on a state machine (like OpenGL), designed to closely resemble that of MATLAB, though its use is discouraged. SciPy makes use of matplotlib.

Matplotlib was originally written by John D. Hunter, has an active development community, and is distributed under a BSD-style license. Michael Droettboom was nominated as matplotlib's lead developer shortly before John Hunter's death in 2012.

4.4 NLTK

The Natural Language Toolkit, or more commonly NLTK, is a suite of libraries and programs for symbolic and statistical natural language processing (NLP) for English written in the Python programming language. NLTK includes graphical demonstrations and sample data.

NLTK is intended to support research and teaching in NLP or closely related areas, including empirical linguistics, cognitive science, artificial intelligence, information retrieval, and machine learning. NLTK has been used successfully as a teaching tool, as an individual study tool, and as a platform for prototyping and building research systems.

4.5 R

R is an open source programming language and software environment for statistical computing and graphics that is supported by the R Foundation for Statistical Computing. The R language is widely used among statisticians and data miners for developing statistical software and data analysis. Polls, surveys of data miners, and studies of scholarly literature databases show that R's popularity has increased substantially in recent years. R and its libraries implement a wide variety of statistical and graphical techniques, including linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, and others. R is easily extensible through functions and extensions, and the R

community is noted for its active contributions in terms of packages.

4.6 Ggplot2

Ggplot2 is a plotting system for R, based on the grammar of graphics, which tries to take the good parts of base and lattice graphics and none of the bad parts. It takes care of many of the fiddly details that make plotting a hassle (like drawing legends) as well as providing a powerful model of graphics that makes it easy to produce complex multi-layered graphics.

4.7 SNA

Social network analysis (SNA) is the process of investigating social structures through the use of networks and graph theory. It characterizes networked structures in terms of nodes (individual actors, people, or things within the network) and the ties, edges, or links (relationships or interactions) that connect them. Examples of social structures commonly visualized through social network analysis include social media networks, memes spread, friendship and acquaintance networks, collaboration graphs, kinship, disease transmission, and sexual relationships. These networks are often visualized through sociograms in which nodes are represented as points and ties are represented as lines.

4.8 Fuzzywuzzy

Fuzzy string matching like a boss. It uses Levenshtein Distance to calculate the differences between sequences in a simple-to-use package.

V. Algorithm

5.1 For performing sentiment analysis

Step 1: Generating the Key for the PRAW (Python Reddit API Wrapper) by means of an existing account on Reddit as a developer.

Step 2: Create an .ini file having the format.

```
[reddit_bot]
Username: reddit username
Password: reddit password
client_id: client_id that was received
client_secret: client_secret that was
received
```

The username is the username used by the user to login and the password is the authentication key for the user. The client_id and the client_secret was generated by Reddit, which allows the bot to scrap its content for analysis.

Step 3: Run the Python Code to fetch the data from the past year (scrap data) and to calculate and plot the data for further analysis.

Step 4: The data set taken can be then used to find the sentiments of users generally as well as for specific topics. It can also be used to find frequency distribution of words appearing in the corpus.

Step 5: Plot the values to have a graphical overview of the analysis that is just the visualization part of the process.

Step 6: Use tools like NLTK to find the positive response, negative or the neutral responses of the users and that analysis in turn depict the nature of the topic.

5.2 Subreddit Visualization using R-

- Extract dataset from reddit using Google BigQuery and the user information to form network graphs. The graph includes the source, destination and the weight.
- The weight is calculated by counting the participation of users in the two subreddits which are above 5 using the comments section.
- The network graph was plotted using R
    A) Network Graph of Reddit Subreddits. (Degree Centrality and Node Type)
    B) Network Graph of Reddit Subreddits. (All Topics Included)
    C) Individual Graph for each Topic.

VI. Methodology

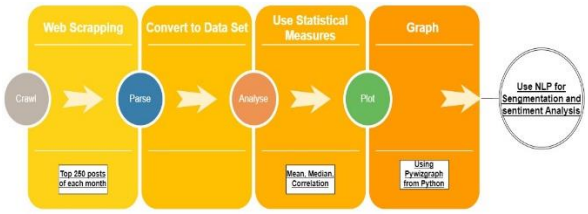


Fig. *Flow of Process in the Sentiment Analysis.*

As shown in the above diagram the Reddit pages would be

1. Scraped using the Bot algorithm designed along with PRAW and Python.
2. Convert the fetched or parsed data into a dataset for ease of analysing.
3. Analyze the data set using software as NLTK used in Natural Language Processing (NLP) and create plots according to the result outputs.
4. Draw conclusion with respect to the analysis about the topic of the Reddit discussion among the users and the community.

VII. Discussions

7.1 Scraping Of Data

Web Scraping or web extraction is a practice employed to extract large quantities of data from websites whereby the data is extracted and saved into a local file in one's computer or to a database in table (spreadsheet) format or excel format.

It includes the use of complex arithmetical algorithms. Screen/web scraping is a method for extracting word-based characters from the web so that they could be analysed. Commonly, it is used to extract characters or paragraphs of interest from the websites (web scraping), though not exclusively.

As the Internet has continued to grow, the amount of data publicly available has snowballed into an unbelievable size. While this data has a lot of power and potential, availability does not necessarily translate into accessibility; due to the structure and immense size of the web, many questions cannot easily be answered using this data. Web scraping provides a great solution to facilitate digesting web data, providing highly specific rules on what data to gather and aggregate. Still, web scraping requires a sophisticated understanding of programming, web technologies such as HTML, and the structure of data on the web (e.g., the Document Object Model (DOM)). As a result, programmers who want to collect data from the web often find themselves writing tedious, complicated scripts and people without technical experience often find themselves unable to collect the data at all.

The first step in our project is to build a simple crawler that scrapes data from Reddit. This will be done using PRAW in Python. To scrape Reddit, we will need to obtain a valid API key from them. The key was first generated on the Reddit profile which gives the access to the crawler bot to crawl the Reddit page legally while binding to all the rules and regulations of the site in order that the sites working is not affected.

7.1.1 Data Scraping and Web Mining of Reddit Using the algorithm Proposed below.

Step 1: Create a reddit account.
Step 2: Navigate to preferences → apps.
Step 3: Get the client id and client secret keys by filling in the details.
Step 4: Create an init file with the following contents-
Reddit username, password, client id, client secret
Step 5: Create an empty file called commented.txt

Step 6: *Authenticate-*
Step 6.1: Read data from init file.
Step 6.2: Build json object and pass data to reddit.
Step 6.3: Authenticate the bot and store the returned data in an object.

Step 7: *Fetch data-*
Step 7.1: Get new comments from the subreddit 'India'.
Step 7.2: If comment text matches a regex of our interested topic,
Step 7.2.1: Save comment id.
Step 7.2.2: Add id to commented text file via file handling.
Step 7.2.3: Save comment body.
Step 7.2.4: Else, go to next comment.

Step 8: *Set schedule-*

Step 8.1: To work around Reddit API limits, set scraping schedule to few minutes.
Step 8.2: Call fetch data
Step 8.3: Wait for few seconds.
Step 8.4: Repeat

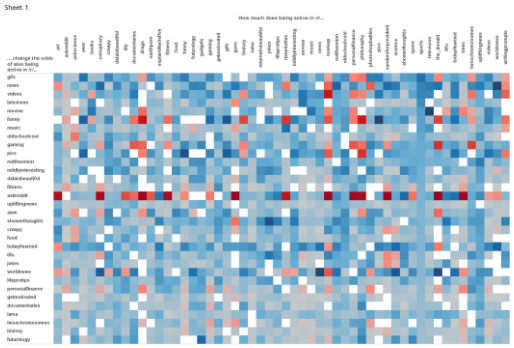
Step 9: Run analysis
Step 9.1: Get total comments parsed.
Step 9.2: Get total comments that are relevant.
Step 9.3: Get percent of relevant comments.
Step 9.4: Run analysis on the data collected.

Step 10: Stop

The above algorithm is useful in scraping the Reddit site and hence the data set created from it can be further used to read and analyze the emotions and the sentiments of the users as a community. This analysis can be used to bifurcate the topics on which the current socio political situations is taken as a positive or negative impact by the masses.

VIII. Results and Observations

8.1 Network Analysis



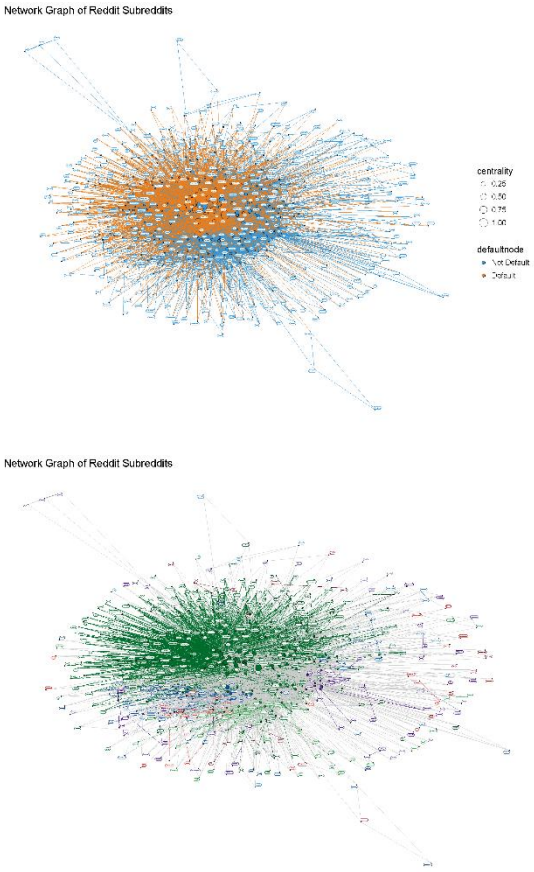
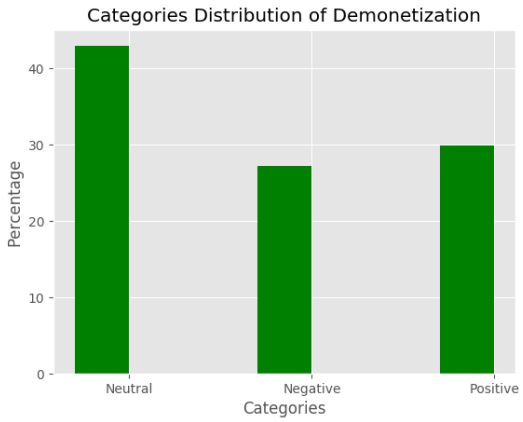


Fig. *showing the correlation table of the likelihood of user posting in different subreddits, the centrality graph, and topic wise breakdown of the centrality graph.*

8.2 Sentiment Analysis



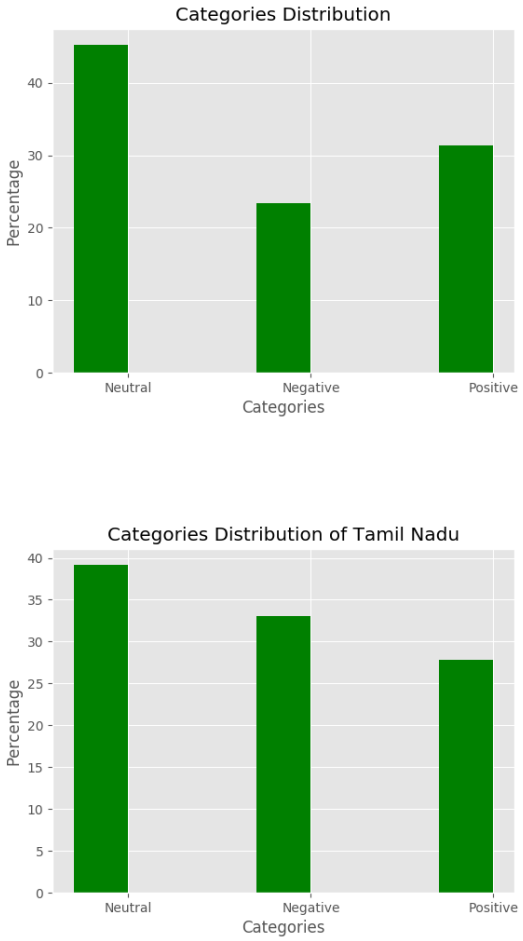


Fig. *showing the fetching the comment from reddit. The first figure is showing sentiment analysis graph for Demonetization, second graph represents the global sentence types and the third one showing analysis about Tamil Nadu.*

Positive comments like good, fine, happy, fun, amazing, lovely, adventurous, advocated, affability, affable, affably, affectation, affection, affectionate, affinity, affirm, affirmation ,affirmative, affluence, affluent, afford, affordable, affordably, affordable, agile, agilely, agility, agreeable, agreeableness, agreeably, all-around, alluring, alluringly, altruistic, altruistically, amaze, amazed etc.

Neutral words like coarse, detached, indifferent, listless, skeptical, serious, solemn, weary etc.

Negative words like abuse, abused, abuses, abusive, abysmal, abysmally, abyss, accidental, accost, accursed, accusation, accusations, accuse, accuses, accusing accusingly, acerbate, acerbic, stall, Stalls, stammer, stampede, standstill, stark, starkly, startle, startling, startlingly, starvation, starve, static, steal, stealing, steals, steep, steeply, stench, stereotype, stereotypical, stereotypically etc.

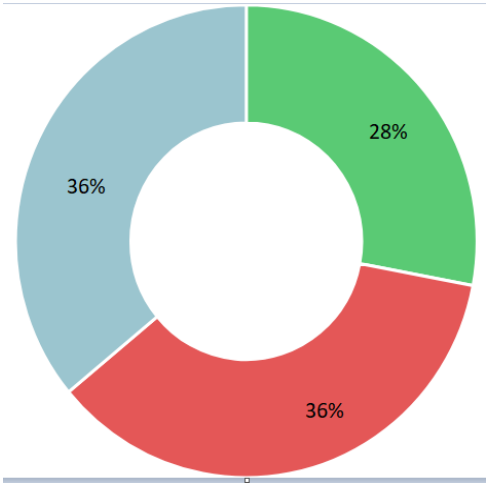


Fig. *Pie chart of result*

The above figure showing result in percentages. 36% percent of comments contain negative and neutral words. Remaining 28% contain positive words.

IX. Summary and Discussion

The first step in my project was to collect huge data to analyze we collected data from Reddit. Crawler start collecting data from popular section on reddit so that data should be from those topics where user are more active and then it crawled through comments, likes and number of votes. Data divided on the basis of date of posting, popularity, highest scoring in a time period. Different users can post topic in different subreddit so we need to list all subreddits relevant to topic. Subreddits' Meta data contains all details accepts comments on topic. Since we have few months to complete project so we take only top 250 posts. That is around 1% of data available on reddit.

After Meta data of list, I collected the comments, submission, and the difference of up and down scores. PRAW library is used which is written in python for crawler. PRAW gives way to share objects of python across the network. All crawler was executed on virtual machines and all of them have access

to shared storage. Data of distinct pages are stored.

One of the difficulty faced in the project is to fight back to blocking system of reddit. Since many crawler wants to collect data from the reddit due to this huge number of request is received by reddit system and system can't handle large number of bot request at time that why it start blocking the crawler. If rate of request of crawler become greater than threshold it will be blocked for some time by anti-crawling technique. IP address of crawler will be black listed. This is very common technique used by this type of web platform. For this crawling speed decrease and the final crawler sent only one request in one minutes.

Reddit normally anticipates creeping in various ways. Like a non-confirmed session cannot peruse the subreddit requested by date. A confirmed client can peruse a subreddit by date, yet the separation you can peruse back is topped. In this manner it is innately unrealistic to get an entire creep of the site. This single factor caused a move in center of the paper at the point when in the exploratory stage because of the failure of acquiring requested accommodation information. One more problem occur that crawler make many attempts to crawls badly or down URLs of reddit because crawler was unable to found difference between down and bad URLs. Sometimes system get restarted because of load and controlling reliability of shared resources. Final result of crawler is in text file that is many times smaller in storing size than the HTML file which saves lots of space and text working with text file is also very easy.

Comments in reddit have a tree like structure and it can be see like a directed acyclic graph or more particularly, a tree with the root that accommodation itself. While each reddit's remark tree is distinct, numerous properties appear to recur. For instance, the higher score a comment has, the more likely it is to have many replies. Many variables can be examined including the tree height, the number of leaf, normal traverse, and many more. These parameters then compared with other socio-Political topics. Using them we can plot graphs, classify data, and clustering for further calculation. The normal comment on Reddit had less than 10 replies, with 95% of all entries having under 100. It has seen more prominent than 5000 comments on a post. Different properties including greatest width and height were found to have comparative appropriations to the general tree size. These qualities shows direst association with the dissemination of the tree.

X. Conclusion

This project provides us a dataset on India and overview of reddit's comment structure. I identified interesting relations and web platform properties among subreddits. This task showed that utilizing NLTK for sentiment analysis on Reddit post and comments is suitable and works fine. This project can give us the insights about the ongoing tends on specific topic to realize the sentiments of people. Application of such methods can be used for marketing, evaluating guests and make operational improvements or capital expenditure.

XI. Future enhancements

This project is limited to reddit because of use of APIs but we can use same algorithm and tools to analyze other web platforms like Twitter and Facebook. Due to reliability of algorithm we used very less data for analysis by improving algorithm we can use this method to analyze large dataset for more accurate result. Some users put Images or Links to other web sites in these case we can't find any useful data to analyze in future we can solve this problem.

References

[1] A Look into the World of Reddit with Neural Networks by Jason Ting

[2] Twitter Sentiment Analysis by Alec Go, Lei Huang, Richa Bhayani

[3] Sentiment Analysis: A Literature Review ZHU Nanli, ZOU Ping, LI Weiguo, CHENG Meng

[4] Twitter Sentiment Analysis: The Good the Bad and the OMG! By Efthymios Kouloumpis, Theresa Wilson, Johanna Moore

[5] http://snap.stanford.edu/class/cs224w2011/proj/tbower_Finalwriteup_v1.pdf
[6] https://cs224d.stanford.edu/reports/TingJason.pdf
[7] https://github.com/jeffreybreen/twitter-sentiment-analysis-tutorial201107/blob/master/data/opinion-lexicon-English/negative-words.txt
[8] https://github.com/jeffreybreen/twitter-sentiment-analysis-tutorial-201107/blob/master/data/opinion-lexicon-English/positive-words.txt
[9] http://sentiwordnet.isti.cnr.it/
[10] https://cs224d.stanford.edu/reports/TingJason.pdf
[11] https://nltk.org/
[12] https://www.reddit.com/r/india/
[13] https://www.lexalytics.com/semantria/excel
[14] https://www.lexalytics.com/technology/sentiment
[15] https://lct-master.org/files/MullenSentimentCourseSlides.pdf
[16] http://www.cs.cornell.edu/home/llee/omsa/omsa-published.pdf
[17] http://jscdev.github.io/Reddit-Sentiment/

Books:

[1] Sentiment Analysis: Mining Opinions, Sentiments, and Emotions

Book by Bing Liu

[2] Opinion Mining and Sentiment Analysis

Book by Bo Pang and Lillian Lee

Appendix

API	An application program interface is code that allows two software programs to communicate with each other
Bot	It is a software application that runs automated tasks (scripts) over the Internet.
Clustering	clustering is the task of grouping a set of objects
Crawling	Collecting Meta data from web.
Creeping	Collecting Meta data from web.
DOM	Document Object Model is an application-programming interface (API) for valid HTML and well-formed XML documents.
Fetching	Downloading web page contains.
Hashing	Hashing is the transformation of a string of characters into a usually shorter fixed-length value or key that represents the original string.
HTML	Hypertext Mark-up Language is the standard mark-up language for creating web pages and web applications.
Interpreter	Interpreter is a computer program that directly executes codes.
Metadata	A set of data that describes and gives information about web page.
OOP	OOP is a programming language model organized around objects rather than "actions" and data rather than logic.
Reliability	The degree to which the result of a measurement, calculation, or specification can be depended on to be accurate.
Sentiment	A view or opinion that is held or expressed.
.ini file	Initialization file is a file extension for an initialization file format used by Microsoft Windows.
NLP	Natural language processing is a field of computer science, artificial intelligence and computational linguistics concerned with the interactions between computers and human languages, pipelines.
Subreddit	A subreddit is like a sub forum on reddit. Most subreddits are focused on one topic.
Session	it is the technique used by the web developer to make the stateless HTTP
XML	Extensible Mark-up Language is a mark-up language that defines a set of rules for encoding documents in a format that is both human-readable and machine-readable